AI Doomsday Bostrom Notes

AI Doomsaying: This is the view that creating superintelligent machines will lead to the destruction of humanity.

The orthogonality thesis: pretty much any amount of intelligence can be combined with pretty much any goal.

> There is nothing about intelligence that would make it so that as something gets smarter its goals align with human goals such as the survival of humanity and other human goals such as benevolence, spirituality, intellectual curiosity, etc.

The instrumental convergence thesis: whatever goal something has, there are other goals we can predict that it will have in order to achieve its main goal.

- survival
- not losing sight of its main goal
- improving its ability to realize its main goal
- perfect the technology it has to achieve its main goal
- acquire resources to achieve its main goal.

The decisive advantage thesis: a superintelligent AI would be able to take over the world.

After all, humans, who are only slightly smarter than other animals, are dominant.

This makes it plausible that super intelligent AI, which is much smarter than humans, could become dominant if that were one of its goals.

The AI Doomsayer's Argument

(1) Bostrom's theses are true.
(2) If Bostrom's theses are true, then AI superintelligence will destroy humanity.
(3) So, AI superintelligence will destroy humanity.

Motivation for (2): Given the orthogonality thesis, superintelligent AI is not guaranteed, merely in virtue of its intelligence, to care about the survival of humanity. We might try to design such AI so that it does care about what we care about. But, AI doomsayer's claim, it is not obvious how we could do that. Like a genie that misses the meaning of a wish because it interprets such wishes excessively literally, rather than as they are intended, it is unclear how we can design AI in such a way that it only does what we want. Given the instrumental convergence thesis, it is very likely that the destruction of humanity would help superintelligent AI achieve its main goal. For the AI will want to continue improving its ability to accomplish its main goal. To do so it will need more and more resources. The conveniently located atoms that make up humans, together with the atoms that make up the resources humans need to survive, could be converted into whatever other form of matter is needed to accomplish the AIs main goal. Finally, given the decisive advantage thesis, it is very likely that we would be unable to stop the AI from converting the matter we need to survive into the matter that is useful to achieve its goal. So once superintelligent AI is created, humanity is probably doomed.

Objection: we could check whether advanced AI is safe for humans by observing its behavior in a controlled environment. If the AI's behavior is friendly to humans, then can we release it confident it will do us no harm. If its behavior is unfriendly, then we can keep it in the controlled environment.

Reply: AI doomsayers point out that both friendly and hostile AI have an incentive to act friendly while contained in a controlled environment. The hostile AI knows it will be kept out of the real world if it doesn't fool us into thinking it is friendly. It will only become hostile to us once we can no longer stop it. And given how smart it is, given all the other advances and benefits AI will have given us by the time we have superintelligent AI, it will likely be able to fool us into thinking that letting it out is just another advance and nothing more.

Objection: we have control over the AI's motives. We can design its motives however we want. So we could just give AI some innocuous goal like make us smile, make us happy, or to make as many paperclips as it can.

Reply: AI doomsayers claim that the AI will find a way to accomplish the goal we have given it that conflicts with our other goals. A shortcut to making me smile for a very powerful AI might be to paralyze my face. A shortcut to making me happy might be to plant pleasure generating electrodes in my brain. And the best way to maximize the number of paperclips is to rearrange all nearby atoms into paperclips thus destroying all humans. In general, AI doomsayers seem to be appealing to the following thought: For any goal we might give a superintelligent AI, the AI may very well find that a best way to accomplish that goal will either

lead to the destruction of humans or, if not that, then some other unexpected and deeply undesirable outcome.